



Online News Popularity

Group 8 Member:

Huang Sixuan A0049228B

Liu Yaowen A0218856R

Yu Zhe A0218820J

Zhang Tongsen A0105567A

Zhang Yixuan A0218975M



Mashable

Mashable is a global, multi-platform media and entertainment company. Powered by its own proprietary technology, Mashable is the go-to source for tech, digital culture and entertainment content for its dedicated and influential audience around the globe.

MONTHLY UNIQUES

45M

SOCIAL MEDIA FOLLOWERS

28M

SHARES/MONTH

7.5M

»»» Exploratory Data Analysis

The dataset includes 39644 news articles published on Mashable in 2013 and 2014.

A total number of 61 features

The goal of the analysis is to predict the popularity of the news articles and find out what contributes to the popularity

Amazon's Streaming Video Library Now a Little Easier to Navigate

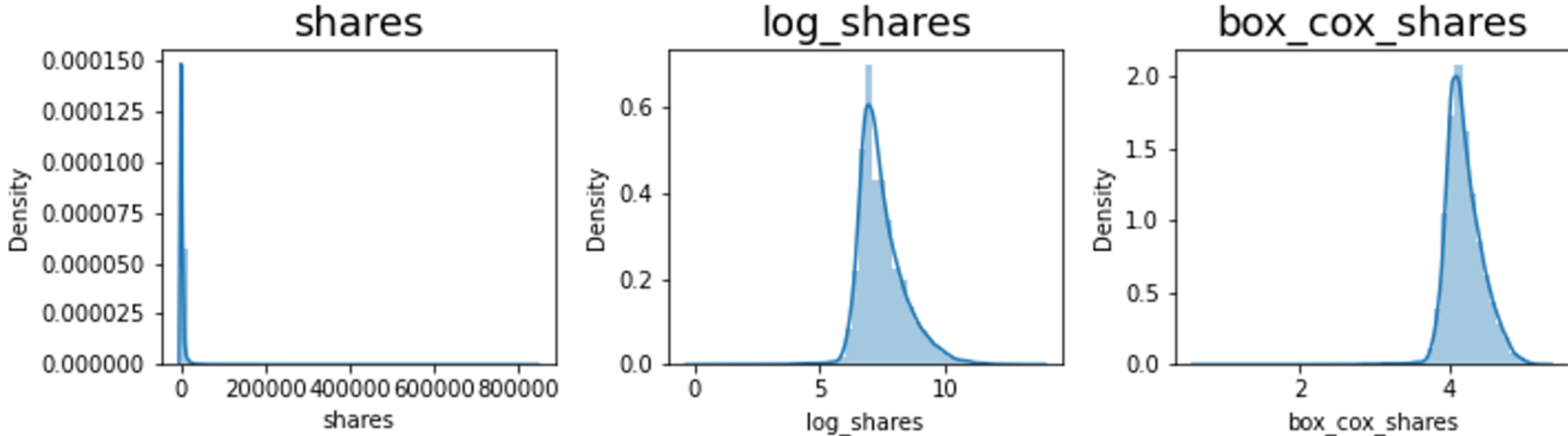


BY LAUREN INDVIK

JAN 08, 2013

Having trouble finding something to watch on [Amazon](#) Instant Video? The retailer launched Monday an [experimental browsing tool](#) that lets users discover movies and TV shows based on their genre preferences or simply the mood they're in.

➤➤➤ Exploratory Data Analysis – Target Variable



Shares Distribution

Shares are mostly small numbers, but there are a lot of extreme large shares and their density is very low. Therefore, the distribution of shares looks very strange. It is difficult to do regression based on original data. We may also need to transform it to classification problem by setting levels of shares.

Log Shares Distribution

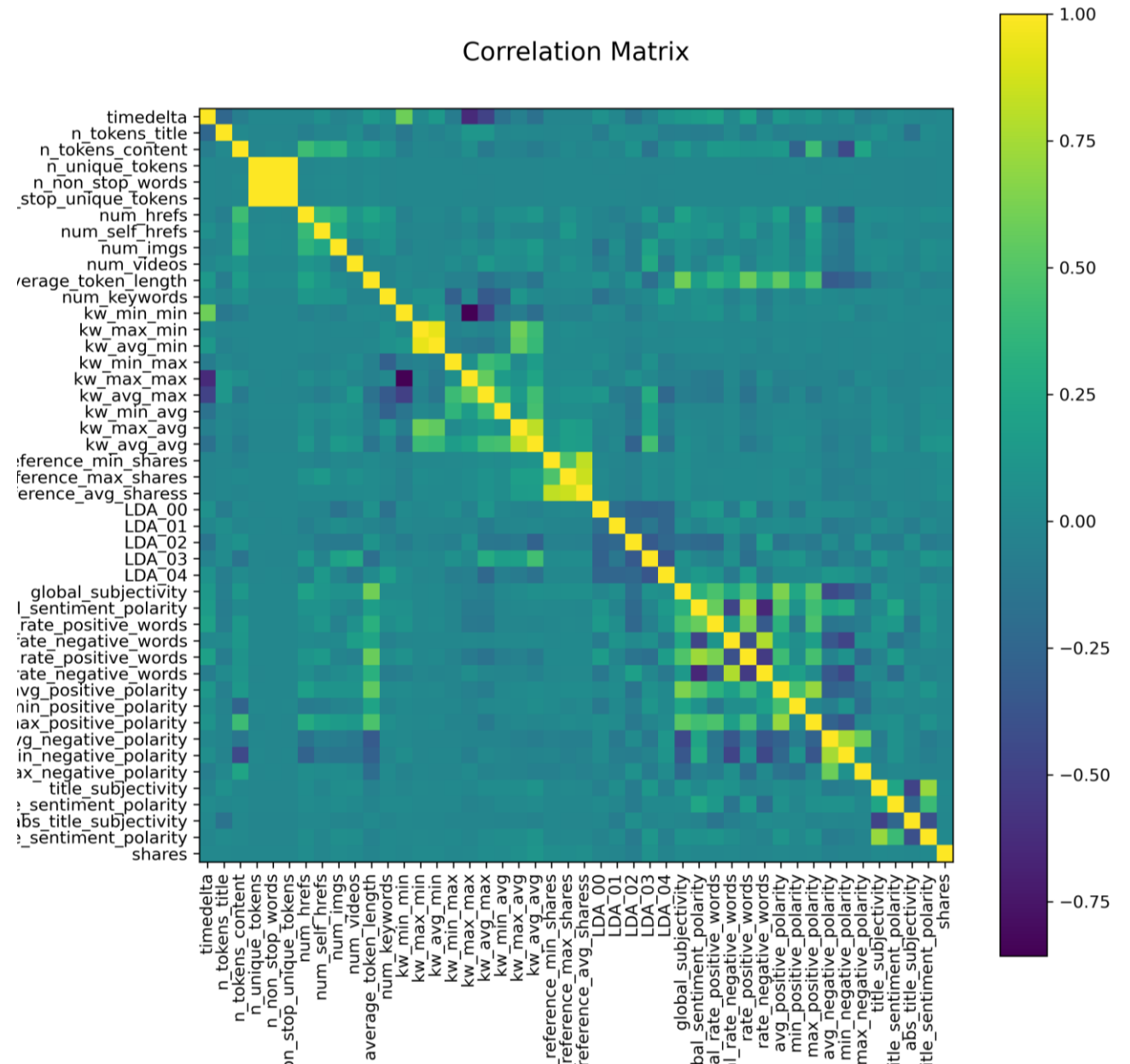
After log processing, the distribution of shares looks like normal distribution. Therefore, in the next stage we may need to use log shares rather than original shares.

Correlation with shares

From the correlation matrix of numerical features, we can see that features have small correlation with shares separately.

Relationships between features

We can also see that there are strong correlation within some features. Therefore, we need to select features rather than use all 60 features.



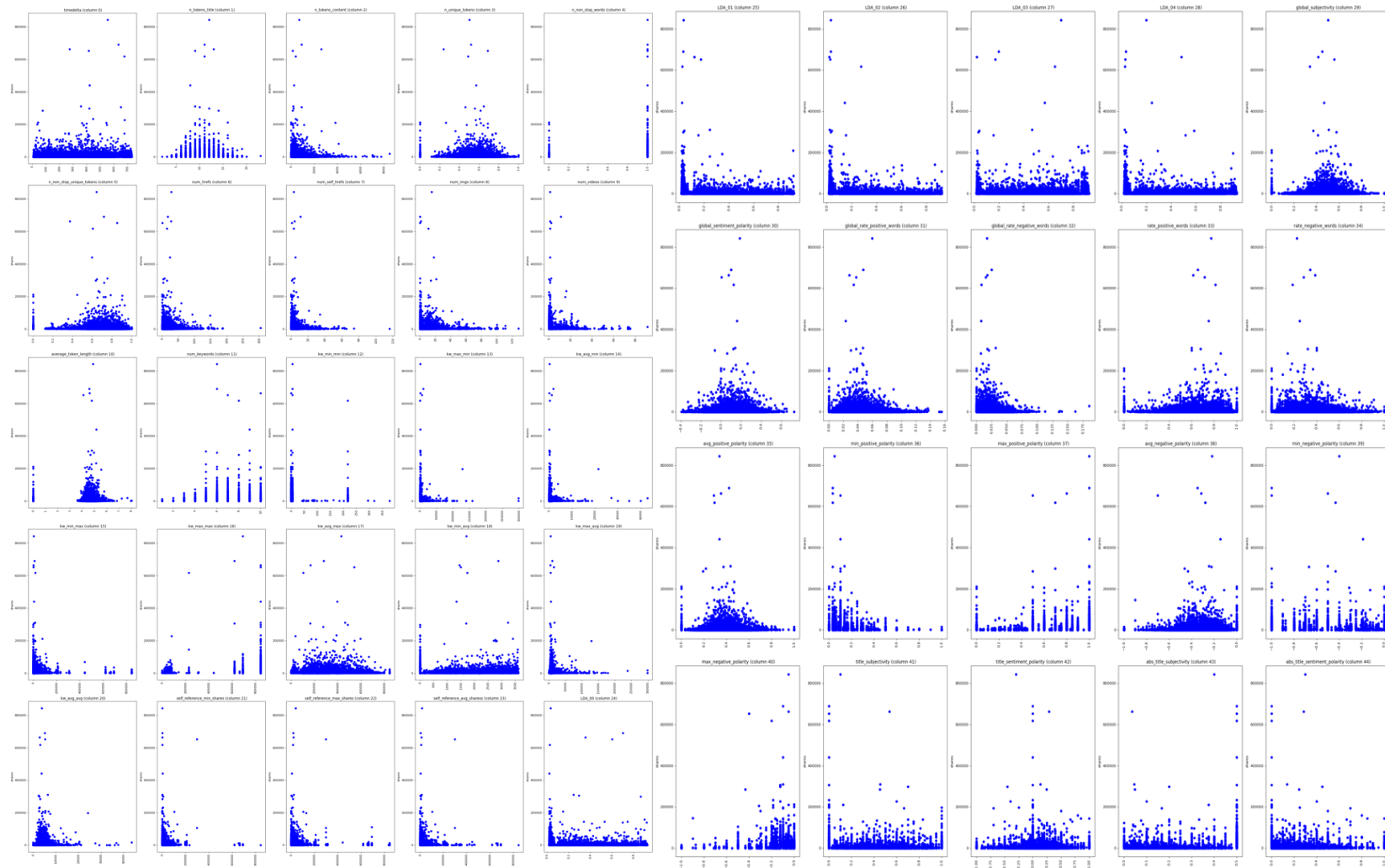
Exploratory Data Analysis – Numerical Features

PCA

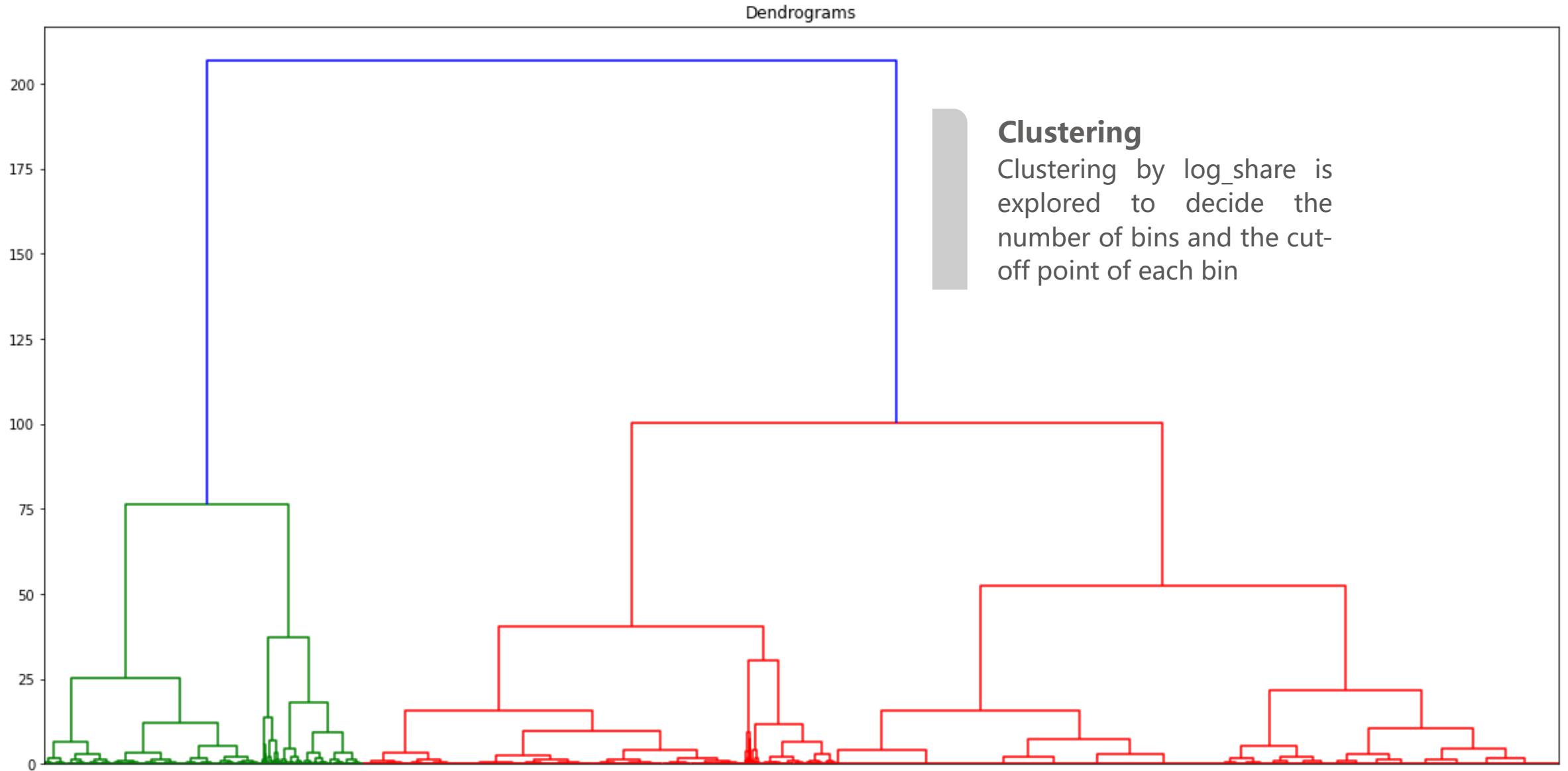
Due to the large number of features, we firstly conducted PCA on numerical features. And the result shows 25 features could represent 99% of all features

Scatter Analysis

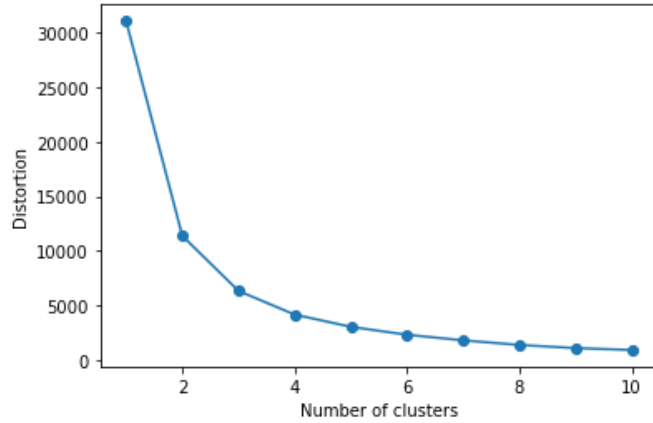
We can also see that there are strong correlation within some features. Therefore, we need to select features rather than use all 60 features.



➤➤➤ Exploratory Data Analysis – Clustering

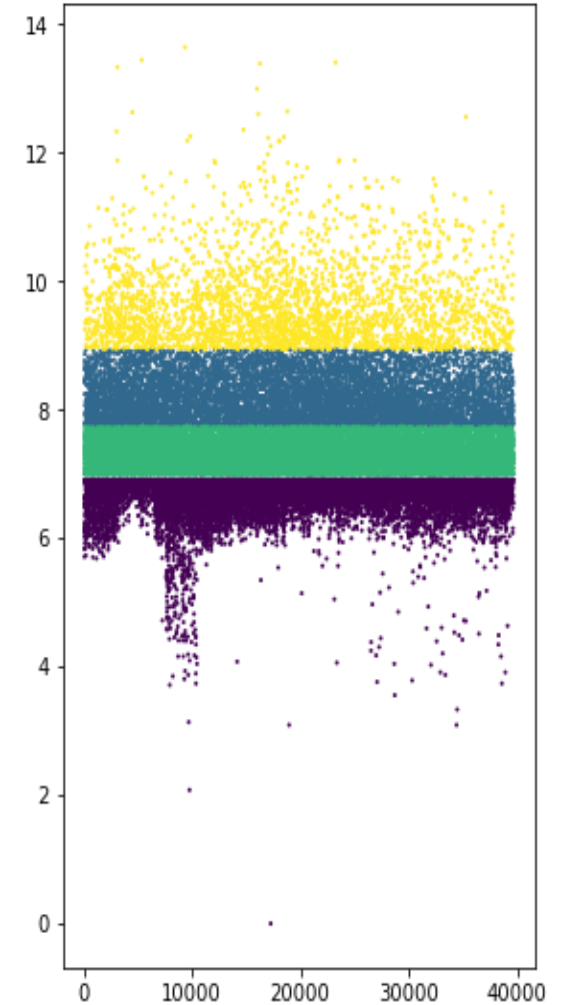
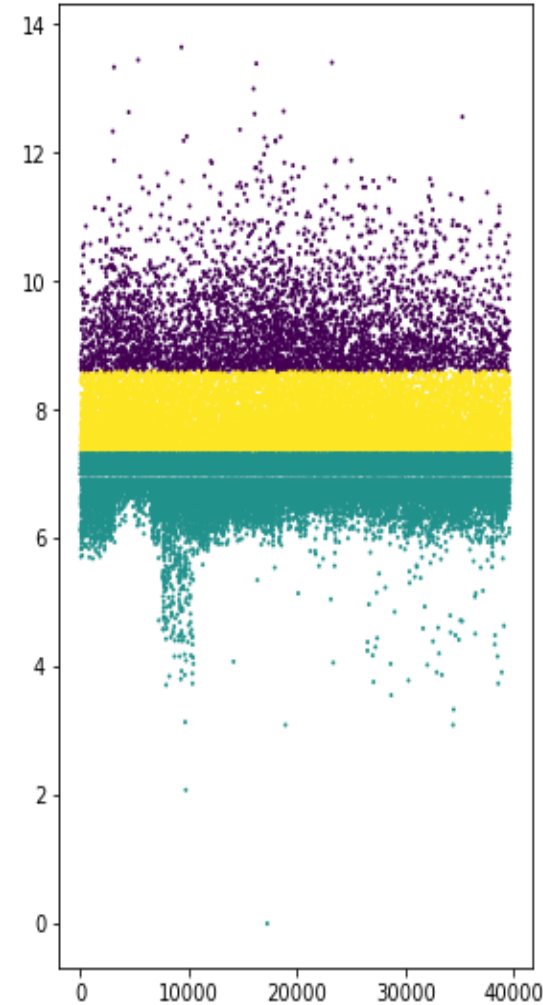
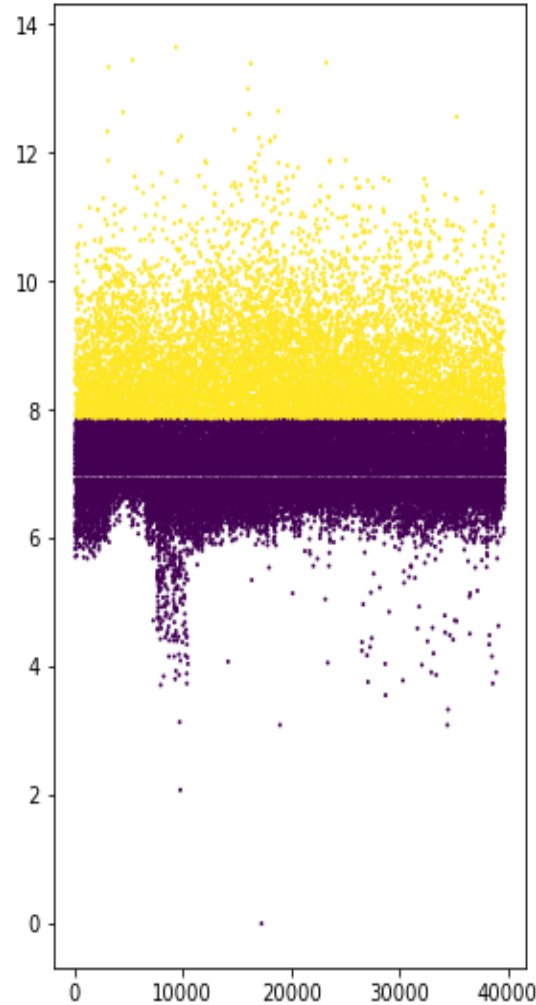


Exp Exploratory Data Analysis – Clustering

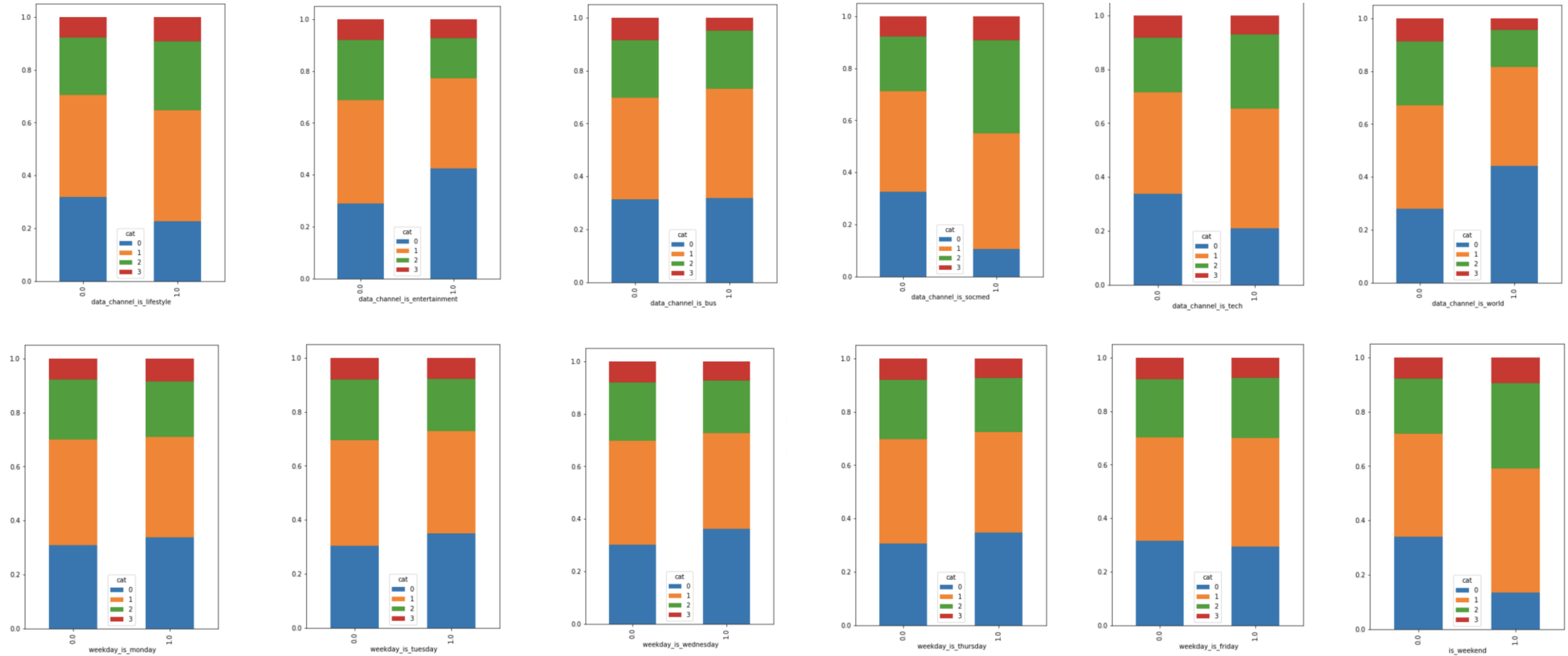


Clustering

Clustering by log_share is explored to decide the number of bins and the cut-off point of each bin



Exploratory Data Analysis – Categorical Features



Channel Features

From the plot we can see that there is little difference on channels between different categories.

Day of week Features

Most of day_of_week variables have similar component in four categories. Only is_weekend, is_tuesday and is_friday look different

Feature Selection

Part I

After EDA and clustering, we firstly chose the features below as input feature to predict the level of news popularity

n_unique_tokens	log_LDA_03
sqrt_self_reference_min_shares	log_LDA_04
sqrt_num_hrefs	global_subjectivity
sqrt_num_self_hrefs	global_sentiment_polarity
average_token_length	global_rate_positive_words
log_kw_max_avg	log_global_rate_negative_words
log_kw_avg_avg	rate_positive_words
weekday_is_tuesday	avg_positive_polarity
weekday_is_friday	avg_negative_polarity
is_weekend	title_subjectivity
log_LDA_00	title_sentiment_polarity
log_LDA_01	all_topic
log_LDA_02	No_word

all_topic

The article does not have a dominant topic across all five LDA topics. We created a binary indicator to show whether the article has all five LDA features lower than 0.3 score or not

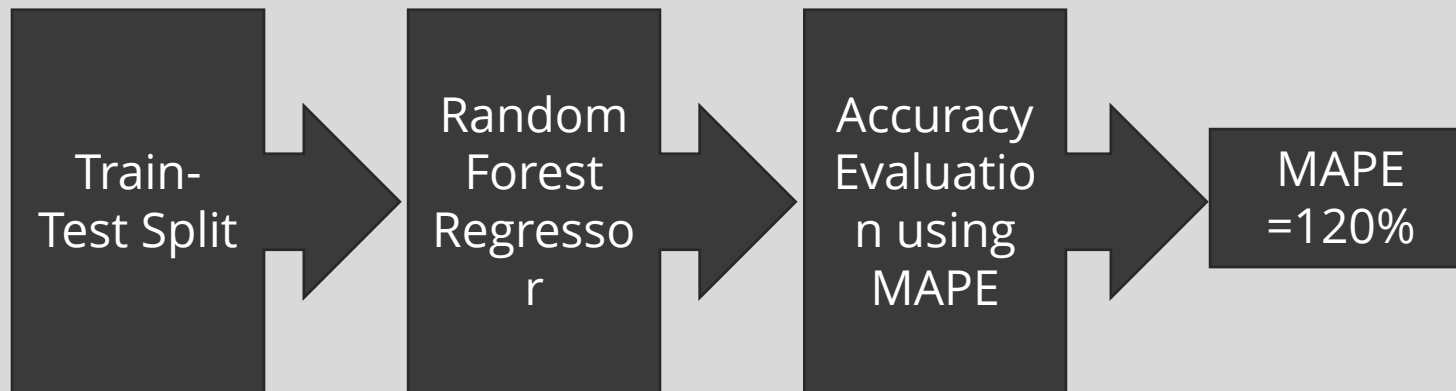
No_word

We think that whether the article contains textual content may be important. Hence, we created a binary indicator to show whether n_tokens_content is zero or not

Model Building

We start by building a regression model to forecast the number of shares for each news article.

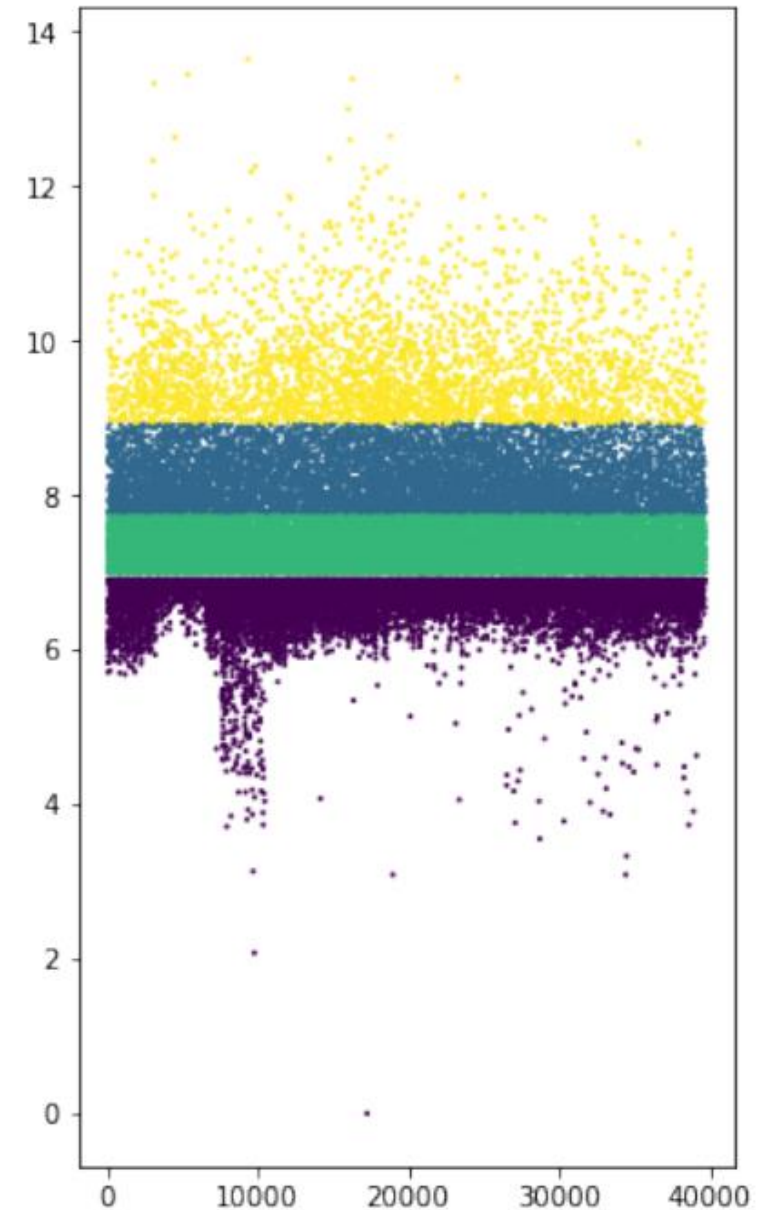
Regression



Model Building

We decided to use 4 clusters for the optimal separation and to make meaningful observations

Clustering



n_unique_tokens	log_LDA_03
sqrt_self_reference_min_shares	log_LDA_04
sqrt_num_hrefs	global_subjectivity
sqrt_num_self_hrefs	global_sentiment_polarity
average_token_length	global_rate_positive_words
log_kw_max_avg	log_global_rate_negative_words
log_kw_avg_avg	rate_positive_words
weekday_is_tuesday	avg_positive_polarity
weekday_is_friday	avg_negative_polarity
is_weekend	title_subjectivity
log_LDA_00	title_sentiment_polarity
log_LDA_01	all_topic
log_LDA_02	No_word

Model Building

Feed the 26 pre-selected features into different classification models

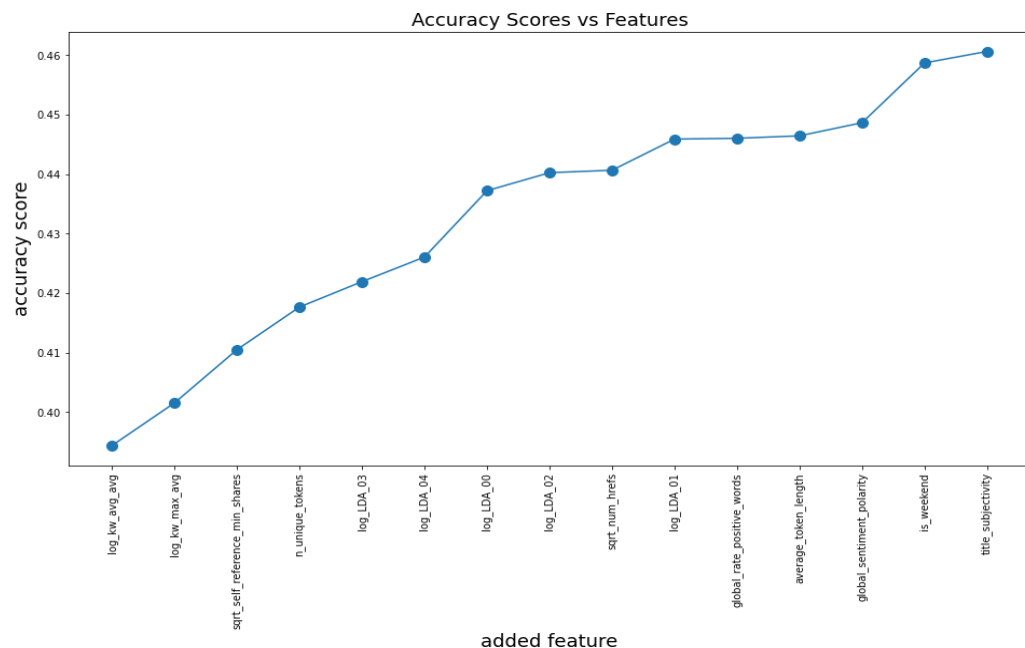
Classification

Fine Tuning

BREA
KING NEWS

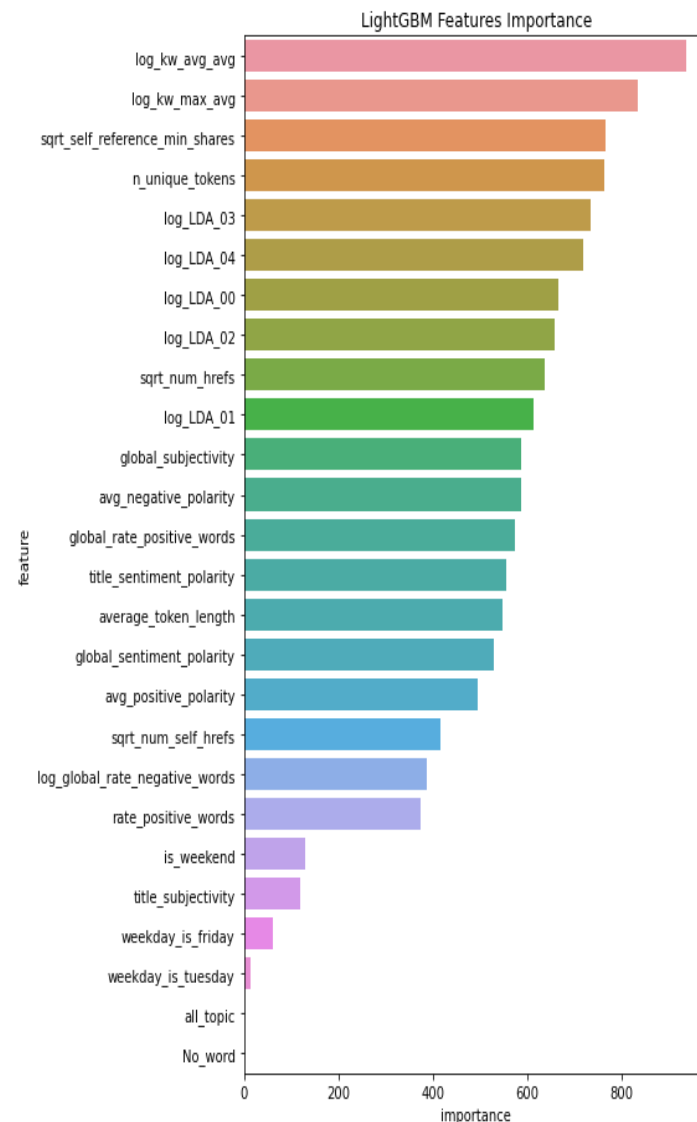
We tried 3 different classification models, used grid search to find their optimal parameters. And for features, we added features into models one by one to find out the optimal input features for each classifier. At the end, we compared the performance matrix of these three classifiers to choose the final model.

Model Building – Light GBM

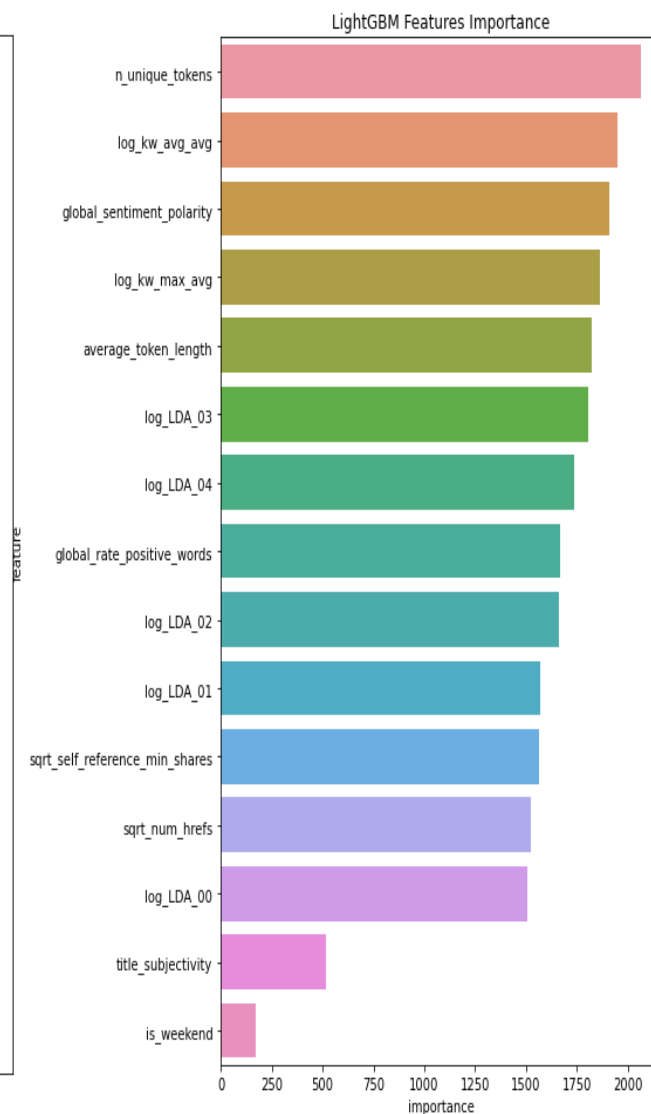


	precision	recall	f1-score	support
0	0.52	0.53	0.52	2252
1	0.44	0.66	0.52	2829
2	0.45	0.18	0.25	1599
3	0.24	0.02	0.04	575
accuracy			0.46	7255
macro avg	0.41	0.35	0.34	7255
weighted avg	0.45	0.46	0.43	7255

auc: 0.6677250300916647
accuracy: 0.4614748449345279



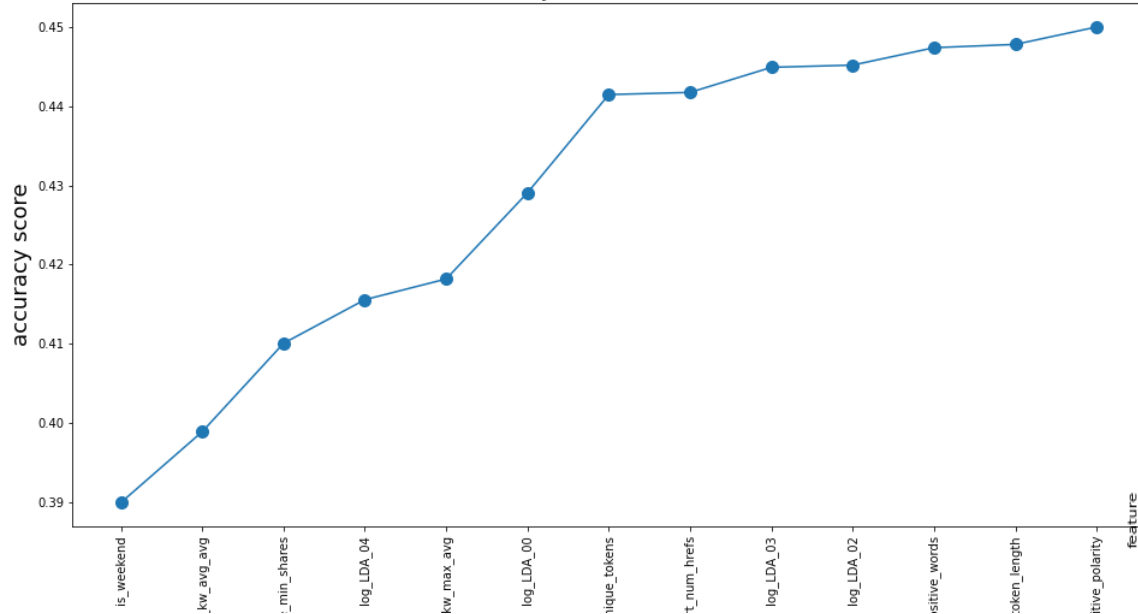
Before selection



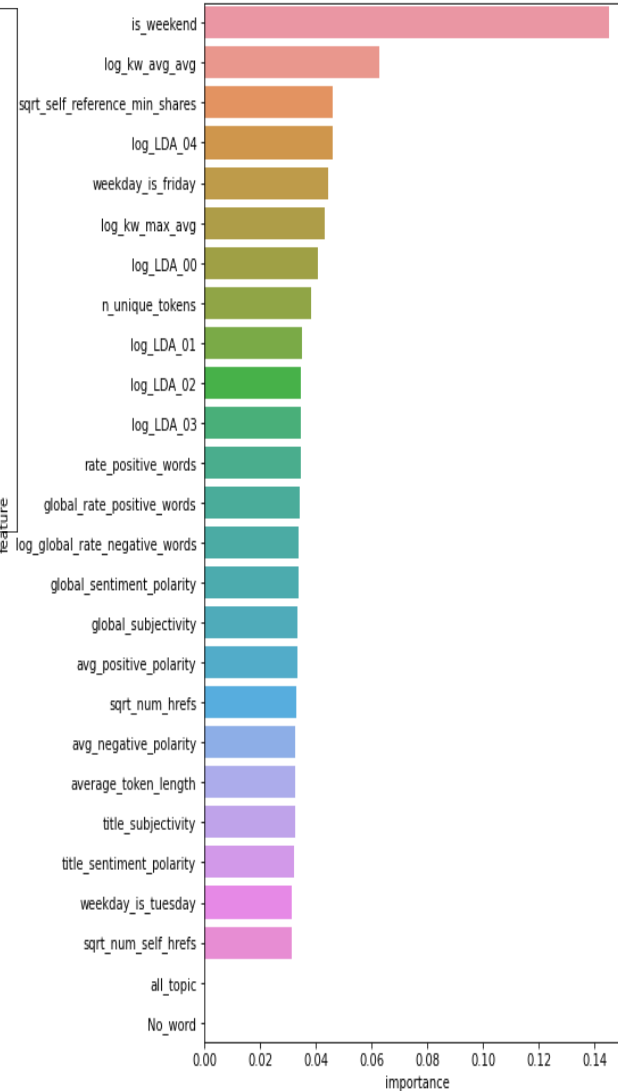
After selection

Model Building - XGBoost

Accuracy Scores vs Features

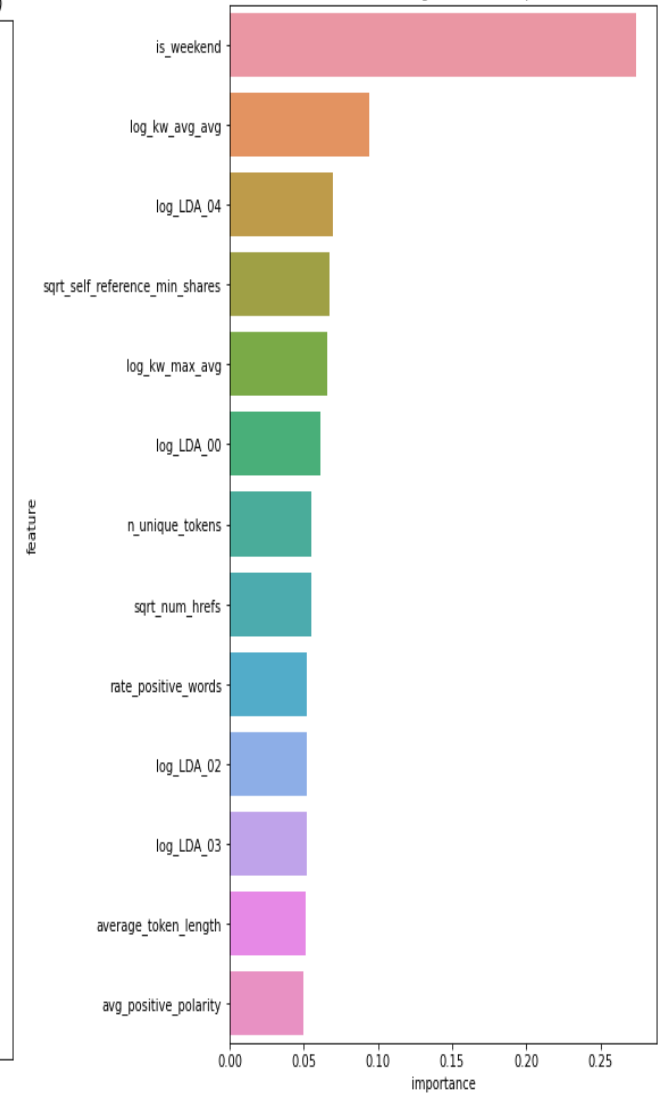


XGB Features 26 features(averaged over store predictions)



Before selection

XGB Features (averaged over store predictions)



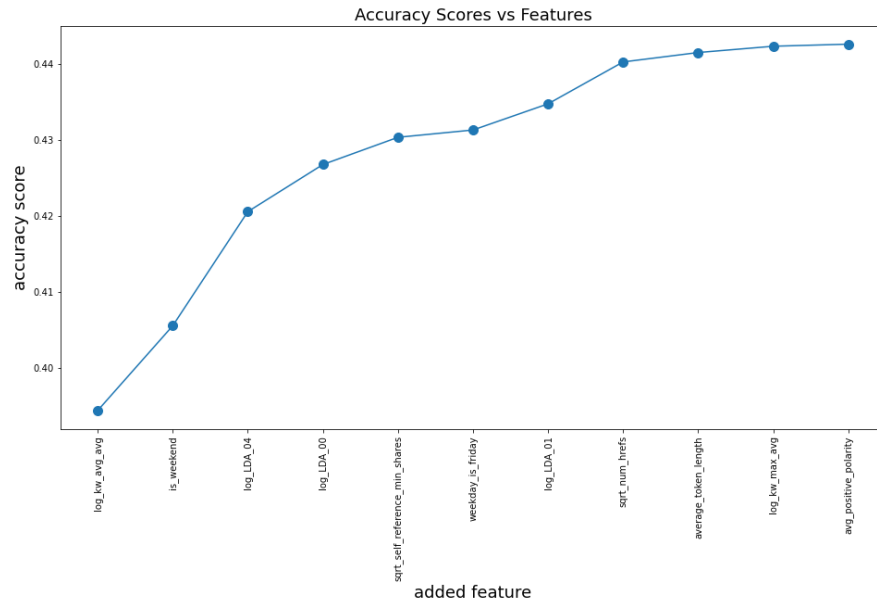
After selection

	precision	recall	f1-score	support
0	0.50	0.53	0.51	2252
1	0.44	0.61	0.51	2829
2	0.40	0.20	0.26	1599
3	0.23	0.05	0.08	575
accuracy			0.45	7255
macro avg	0.39	0.35	0.34	7255
weighted avg	0.43	0.45	0.42	7255

auc: 0.6627588034546997

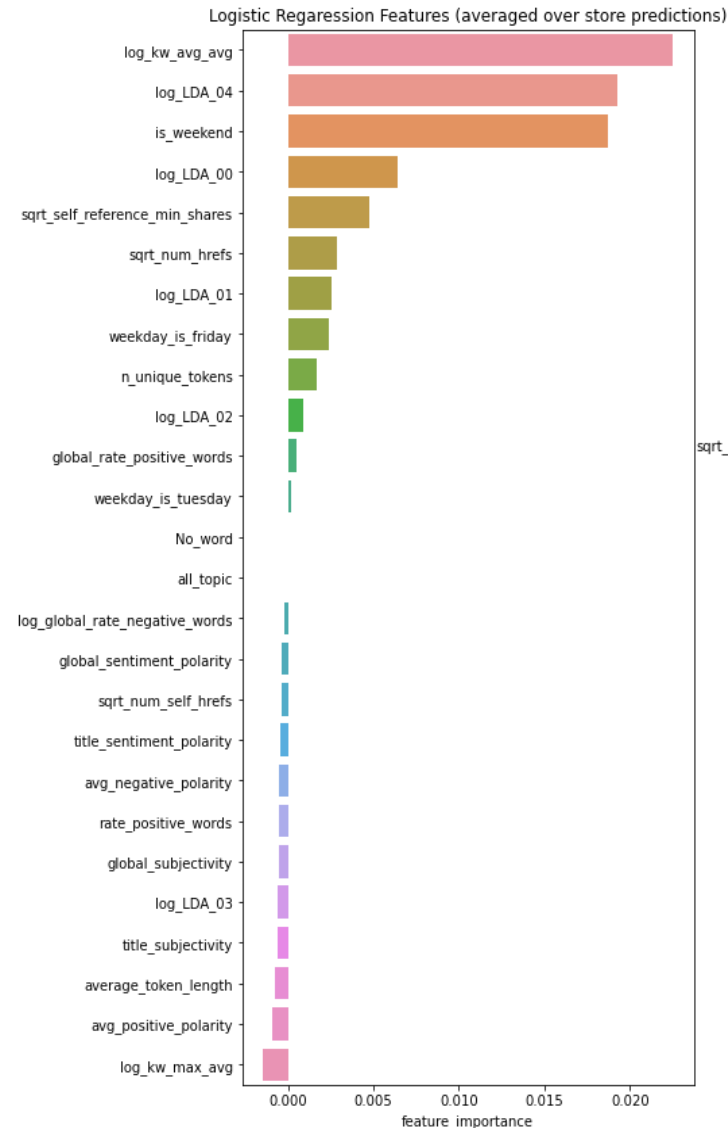
accuracy: 0.4500344589937974

Model Building – Logistic Regression

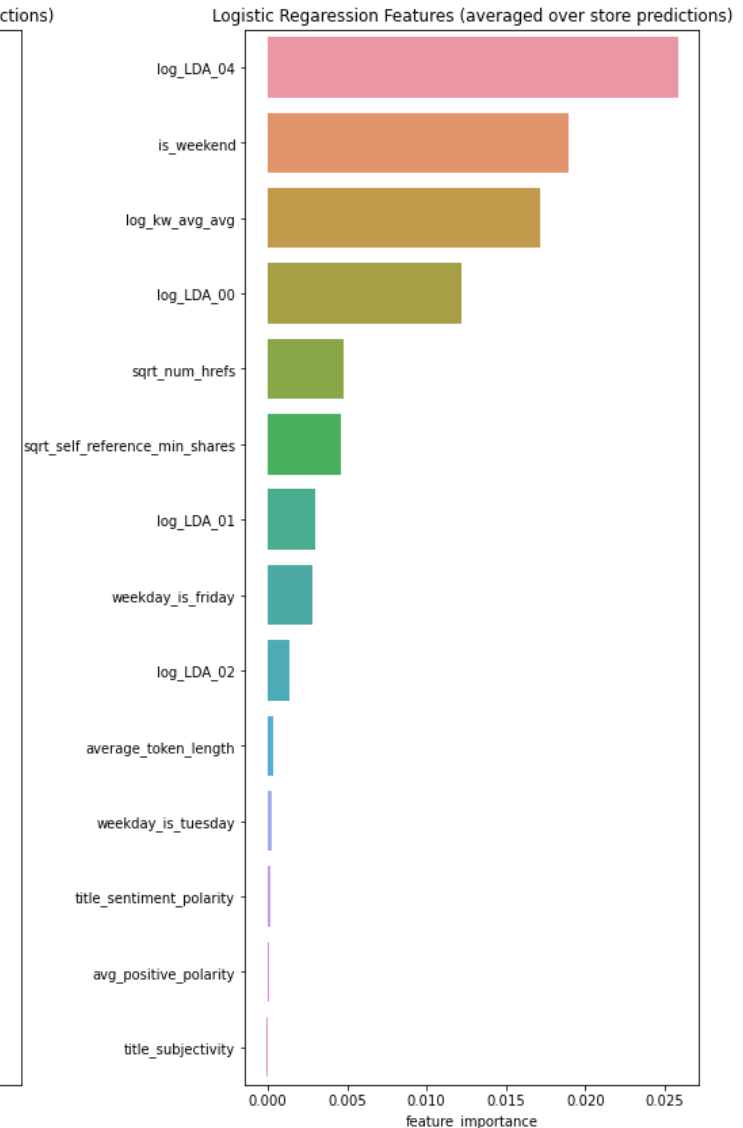


	precision	recall	f1-score	support
0	0.50	0.49	0.49	2252
1	0.42	0.71	0.53	2829
2	0.42	0.07	0.12	1599
3	0.36	0.02	0.04	575
accuracy			0.45	7255
macro avg	0.43	0.32	0.30	7255
weighted avg	0.44	0.45	0.39	7255

roc_auc_score 0.6511766823644244
accuracy_score 0.4460372157133012



Before selection



After selection

Final Model & Interpretation

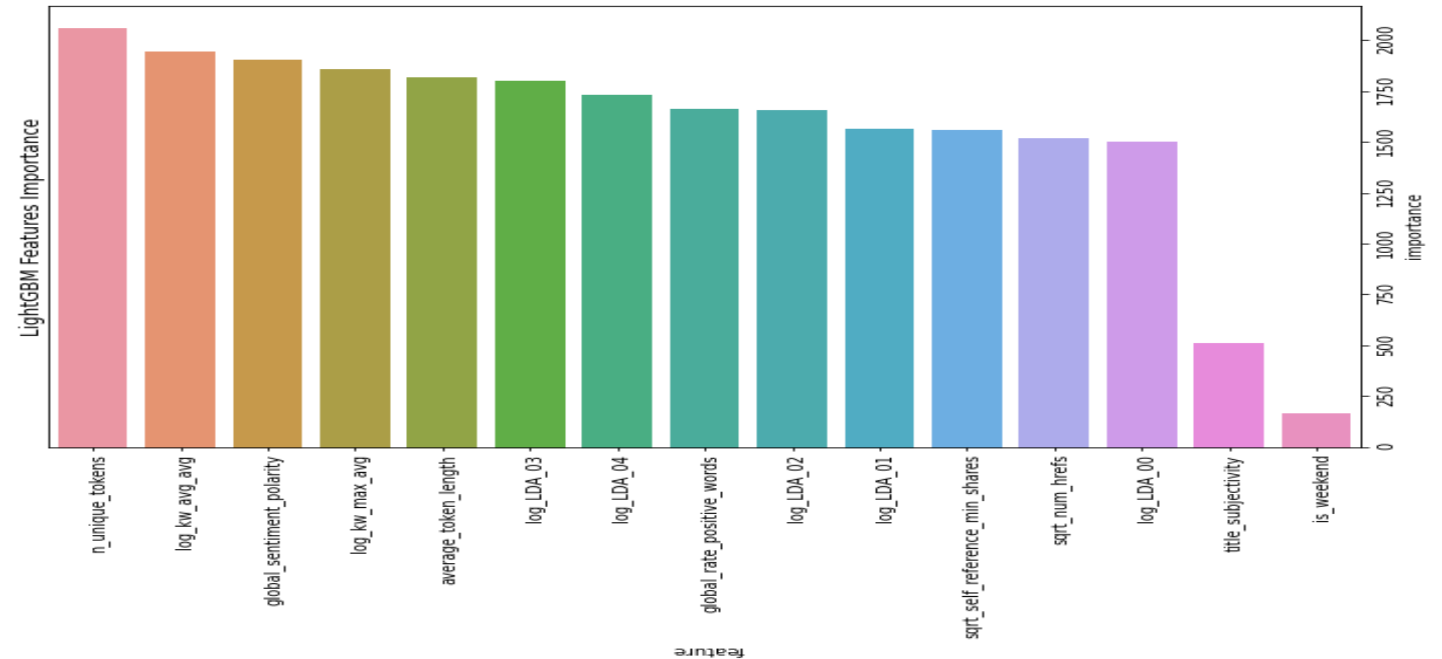
Final Model

This table gives performances of 3 models after tuning parameters. Comparing to other two models, Light GBM has better performance on prediction. Therefore, we choose Light GBM as our final model.


Model	Accuracy	F1-Score	ROC-AUC
Light GBM	0.46	0.43	0.67
XGBoost	0.45	0.42	0.66
Logistic Regression	0.45	0.39	0.65

Final Selected Features

At the end, we used 15 of 26 features to train Light GBM model. The combining use of these 15 features gives best prediction accuracy and it also reduce the number of features we need to consider on predicting popularity of news.



How to write Popular Online News



Increase the number of key words in the article to grab readers' attention



Make the news topic prominent, easier to attract interests



News articles published on weekend tend to have higher popularity



We carefully selected 15 features to conclude the information we can gather from the dataset, chose the best performing model to predict the classification and drew insights from the import features

THANK YOU

Group 8